Published in: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, September 9-13, 2012. International Speech Communications Association, 2012. pp. 2418-2421.

More on the Normalization of Syllable Prominence Ratings

Christopher Sappok[1], *Denis Arnold*[2]

[1]German Studies, University of Duisburg-Essen, Germany
[2]Quantitative Linguistics, University of Tübingen, Germany
christopher.sappok@uni-due.de, denis.arnold@uni-tuebingen.de

Abstract

The perception of syllable prominence depends to a limited extent on the acoustic properties of the speech signal in question. Psychoacoustic factors are involved as well. Thus, research often relies on two types of data: subjective prominence ratings collected in perception experiments and acoustic measures. A problem with the rating data is noise resulting from individual approaches to the rating task. This paper addresses the question of how this noise can be reduced by normalization, evaluating 12 normalization methods. In a perception experiment, prominence ratings concerning German read speech were collected. From the raw rating data 12 different 'mirror' data-sets were computed according to the 12 methods. Each mirror data-set was correlated with the same set of underlying acoustic data. The multiple regression setup included raw syllable duration as well as within-syllable maximum F0 and intensity. Adjusted r^2-values could be raised considerably with selected methods.

Index Terms: perception experiment, inter-rater variability, intra-rater variability, read speech, German, prose, poetry

1. Introduction

This paper is a follow-up to [1], where we presented a number of normalization methods for syllable prominence ratings and preliminary results of an evaluation thereof. Details are summarized in Section 2. In Section 3 of the present paper, additional data and methodological improvements of the original analytical setup [1] are introduced. Section 4 covers the analyses carried out and presents the outcome of the evaluation. The two most successful methods are discussed. Section 5 gives an outlook and illustrates the actual profit from the most successful normalization with specific cases.

The idea behind the normalization of rating data is to reduce what has been called in the abstract: 'noise resulting from individual approaches to the rating task.' This is best explained by taking the view of one of the listeners in our perception experiment, as she goes through eight rounds of the following:

On clicking a button, a signal with an utterance (in German, 30 phonetic syllables) sounds automatically one time. The signal can be replayed without limit, always sounding as a whole. The screen displays 30 vertical slide controls. On clicking anywhere on each slide control, a slider appears that can be moved and moved again in any order. This is how the listener indicates how salient the syllable in question is, according to her hearing-impression. The sliders are scaled 0-30 ([2], Figure 1).

The set of 30 measures collected in one of these rounds is referred to as one 'rating' of one 'signal' here. Each of the 8 signals is an utterance of the same 30 phonological syllables uttered by a different male speaker, durations ranging from 5 to 11s due to differences in speaking rate, phrasing/pausing etc.

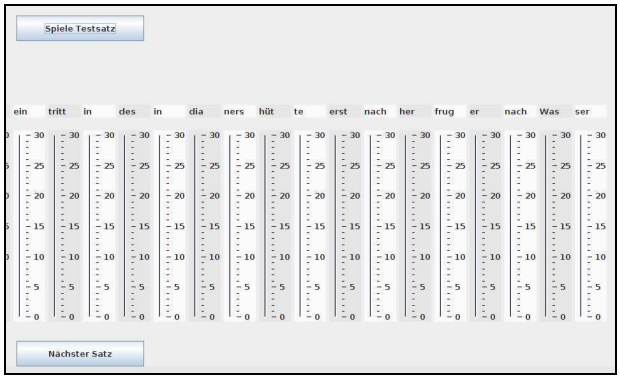

Figure 1: *Right part of a screenshot of the arrangement used to rate one signal (covering 17 out of 30 syllables; "Play"-button above, "Next"-button below).*

Our perception experiment involved 56 listeners rating 32 signals by 8 speakers, delivering N = 13,440 single rating measures (per speaker: M = 1,660). It is most likely that the circumstances described above caused a considerable amount of inter- and intra-listener variability (= 'noise'). As to the specific nature of this noise, we developed the following hypotheses:

1. Listeners project an imaginary horizontal base-line onto the arrangement of slide controls and display *inter-listener* differences regarding the exact positions of these base-lines.
2. Base-line positions are also prone to *intra-listener* shifting in the course of one listener going through signal 1 to 8.
3. Resulting noise can be reduced by setting central tendency = 0 with reference to the sample distribution of one rating of one signal (n = 30).
4. As to the specific parameter, the median is more suitable than the mean, because each base-line thus manifests itself in the form of straight zeroes.
5. Listeners differ in terms of rating-"generosity". Here, resulting noise cannot be met by setting dispersion = 1 with reference to the intra-rating sample distribution as above, because this would imply that the signals themselves do not differ in terms of broadness of prominence variation. Thus, the appropriate reference sample distribution for dispersion normalization is the total amount of rating measures delivered by one specific listener (m = 8*n = 240).

For the evaluation, the following heuristic was introduced in [1]: The more successful noise would be reduced, the higher the normalized rating data would correlate with acoustic prominence in terms of raw syllable duration. Therefore, for each mirror data-set Pearson product-moment correlation coefficients were computed with respect to one and the same set of syllable duration measures.

The results strongly indicated that two methods were equally superior to the rest while differing from each other in terms of computing effort: With 'method 3', intra-rating mean was subtracted from each individual measure (see above: hypothesis 3). Dispersion normalization was not carried out at all. With 'method 5', each individual measure was divided by intra-listener mean absolute deviation (see above: hypothesis 5). The fact that method 3 and method 5 yielded the same level of success (details in Section 4) led to the conclusion that dispersion normalization may be unnecessary [1].

The main objective of the present paper is to either confirm or question this conclusion. To do so, 12 out of the 16 original methods were re-evaluated (the remainig 4 were abandoned for conceptual reasons, see section 2.4.), this time including F0- and intensity measures in a multiple regression setup. The eventual dependent variable in the evaluation of methods is goodness of fit in terms of adjusted r^2 as prompted by R [3] on command:

```
lm(RATING ~ DUR + F0 + INT)          (1)
```

2. Data Acquisition and Preparation

The present section summarizes the acquisition of the "Gold"-corpus (960 phonetic syllables), from which acoustic and perceptual data were derived. Then it describes how each phonetic syllable was associated with a primitive duration, F0, and intensity measure. Finally, it is shown how each phonetic syllable was associated with 14 syllable specific rating measures.

2.1. The "Gold"-Corpus

The "Gold"-corpus consists of 32 readings of the same wording: 8 male German speakers read this wording 4 times, along with a context in which it was embedded. The complete text (wording + context = 123 syllables) originates from the epic poem "Bimini" [4] by German poet Heinrich Heine (1797-1857). It consists of 4 rhymeless stanzas à 4 verses à 4 trochaic feet. The wording in question is the third of the four stanzas mentioned (in the following referred to as 'stanza 3'):

Gold war jetzt das erste Wort, (2)
Das der Spanier sprach beim Eintritt
In des Indianers Hütte -
Erst nachher frug er nach Wasser.

It may be translated into English fairly well preserving word order and meter:

Gold was now the prim'ry word (3)
That the Spaniard spoke on ent'ring
In the Native Indian's shelter -
Only then ask'd he for water.

As a stimulus, the text was presented in two ways, once in the original 'lyrical' layout (LYR) and once transformed and in prose layout (PROS): In stanzas 1, 2 and 4, word order was changed (preserving syntactic structure) in order to spoil the balanced metric organization of the original. Line breaks were deleted, leaving the line-organization of the entire text to purely length-of-string based word processing. The reason why this particular text was selected is the meter underlying the first 5 syllables of the 4th verse of stanza 3. We hypothesized that under layout condition PROS they would preferably be read iambically (4), whereas under layout condition LYR they would – as a priming effect – preferably be read trochaically (5):

erst NACHher FRUG er nach WASser (4)

ERST nachHER frug ER nach WASser (5)

Figure 2 illustrates the raw rating measures with data collected in connection with one reader reading this part first as PROS and then as LYR. For the first three syllables the layout condition tends to have the predicted effect (especially evident in the third syllable; see also Figure 7).

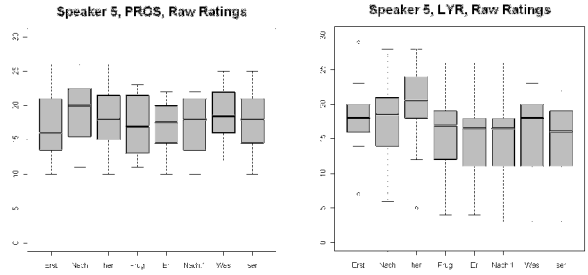

Figure 2: *Boxplots of 14 raw prominence ratings per phonetic syllable concerning two readings of the last sentence of stanza 3 (condition PROS associated with (4), condition LYR associated with (5)).*

In order to elicit maximally different prominence distributions, other conditions were controlled besides the layout condition, the most influential one being that speakers 1 to 4 were laymen and speakers 5 to 8 were university professors of rhetoric (see below: Figure 6). The sessions were recorded with the DAT-recorder SONY TCD-D100 and the SONY ECM-T140 microphone in mono at a sampling rate of 48 kHz. Then, the stanza-3-passages (= 'signals') were extracted and sampled down to 16 kHz.

2.2. Acoustic Measures

The 32 signals were labeled on a segmental level by the first author and independently by an assistant (graduate student), using PRAAT [5] and following the 'liberal phonemization'-standard of SAMPA-D-VMlex V1.0 [6], additionally documenting boundary phenomena (pauses, pre-pause lengthening and lengthening without adjacent pause). After comparison, further steps were based on the first author's segment-boundaries. Duration was measured from the beginning of one onset-initial segment to the next (DUR). In a subsequent step, the assistant derived within-syllable maximum F0 and within-syllable maximum intensity (INT). Then all DUR, F0 and INT measures of pre-pause syllables and syllables lengthened without adjacent pause were deleted manually, because their durations are confounded with boundary phenomena. 161 out of 960 measures were affected. The following analyses refer to the remaining 799 phonetic syllables of fluent speech.

2.3. Perception Measures

The perception experiment originally involved 64 listeners (students), each rating 8 out of the 32 signals of the "Gold"-corpus [1]. Beforehand, the corpus had been split up into 4 packages à 8 signals at random, except that each package contained one signal of each speaker, thus making sure that every speaker was covered by all listeners to the same extent.

Every listener was assigned one package at random, except that each package was treated 64:4 = 16 times. Within the individual listening sessions, the order of the 8 signals in question was randomized again, without restriction. A screening of the data indicated that some listeners had been uncooperative [1] and the ratings of the two most uncooperative listeners per package were discarded. Thus, every phonetic syllable of the corpus was associated with 14 prominence ratings eventually.

2.4. Normalization Methods

The amount of 12 normalization methods is the result of a hierarchy of 3 factors systematically taken into account and the a posteriori discarding of those methods in which only dispersion was normalized, because with dispersion set to 1, inter-rating differences in unnormalized central tendency resulted in great scaling differences and very little or no correlation at all [1].

CT	MEAN								MEDIAN							
CTRD	NONE			IRAT			ILIST		NONE			IRAT			ILIST			
DISPRD	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST
methods		*1*	*2*	**3**	*4*	**5**	*6*	*7*	*8*		*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*

Figure 3: *The factors (grey) and their levels (white) underlying the 12 normalization methods dealt with here (crossed fields represent raw data, slashed fields fields represent very ineffective methods; numbers 3 and 5 indicate the methods identified as most effective in [1]).*

The first factor, CT, governs what the normalization of central tendency is based on, MEAN vs. MEDIAN.

The second factor, CTRD, governs the central tendency reference distribution: absence of CT-normalization (NONE), intra-rating reference distribution (IRAT, n = 30), and intra-listener reference distribution (ILIST, m = 240).

The third factor, DISPRD, governs dispersion reference distribution (the dispersion parameter employed is mean absolute deviation from CT with MEAN as well as MEDIAN, since sd does not work well with MEDIAN): absence of dispersion-normalization (NONE), intra-rating reference distribution (IRAT), and intra-listener reference distribution (ILIST).

Three technical problems with respect to the evaluative analyses remain to be pointed out and clarified:

1. 'Sample-size-difference': Each phonetic syllable of the corpus is associated with 14 individual rating measures on the one hand and three acoustic measures on the other. Thus, all analyses were carried out twofold per method: once the acoustic measures were copied 14 times to fill slots with every rating measure (ALL) and once the rating measures were averaged per phonetic syllable to fill just one slot in line with the three acoustic measures (SMALL).
2. Speaker-differences in 'acoustic behavior' (individual base-lines concerning, e.g., speaking rate, pitch, sonority) result in restricted commensurability. Therefore, we analyzed the entire set of normalized single ratings (TOGETHER, N = 13,440) and additionally the 8 speaker-specific subsets per method (SPEAKER-SPECIFIC, M = 1,660 per subset).
3. 'Commensurability of results': In [1], several derivates of Pearson's correlation coefficients were employed. For clarity's sake, in the present paper the eventual dependent variable is explained variance (adjusted r^2) with all analyses.

3. Results

To describe the raw, unnormalized rating data, Figure 4 shows a histogram and relations to the acoustic measures. The mean of the ratings is 16.75, sd is 5.63. The right side of Figure 4 shows the outcomes of simple regression with respect to single acoustic parameters as well as multiple regression including all three parameters. No correlation was found concerning F0.

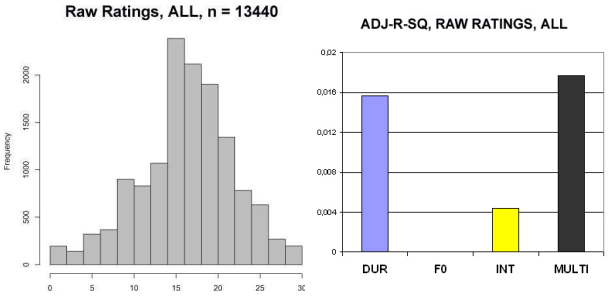

Figure 4: *Raw rating data, left: histogram, right: adj. r^2-values for simple (DUR, F0, INT) and multiple (MULTI) regression analyses; data: ALL, TOGETHER.*

Figure 5 shows adjusted r^2-values from the parameter-specific single and from the multiple regression analyses.

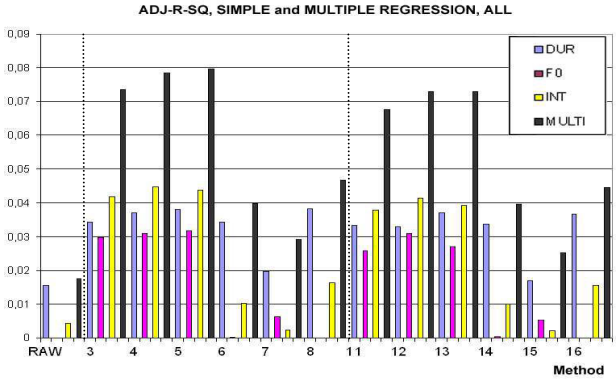

Figure 5: *Single and multiple regression adj. r^2-values for raw data and data normalized according to methods 3 to 8 (mean-based) and 11 to 16 (median-based), separated by dotted lines; data: ALL, TOGETHER.*

Most evident is that normalization could unveil F0-correlation. INT-correlation rises above DUR-correlation in comparison to the raw data with most methods. Generally, normalization affects the three parameters proportionally: the ranking INT > DUR > F0 is maintained in almost every case. The mean-based methods (3 to 8) show a similar profile compared to the median-based methods (11 to 16), being definitely higher (in contradiction to hypothesis 4, see section 1).

Concerning the question of the most effective method, the tendency found in [1] based on just the duration data correlation is confirmed: It all boils down to method 3 vs. method 5 (as method 4 is forbidden due to conceptual considerations, see section 1, hypothesis 5). In order to shed more light on the differences between method 3 and method 5, Figure 6 shows the simple and multiple adj. r^2-values for speaker-specific models with respect to these two methods.

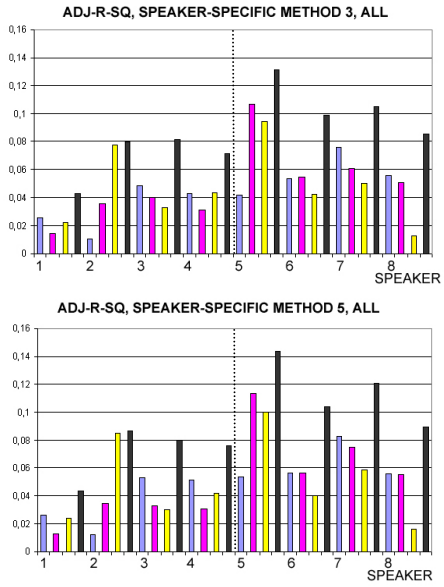

Figure 6: *Method 3 vs. method 5: simple and multiple regression adj. r^2-values from speaker-specific models (speakers 1 to 4: laymen, 5 to 8: professionals, separated by dotted lines); data: ALL, SPEAKER-SPECIFIC.*

Very similar overall profiles can be observed (note the consistent difference for laymen vs. professionals), with most values slightly higher for method 5 (ranging from the outcomes for speaker 3 to the outcomes for speaker 7) . Similar tendencies in [1] were not sound enough to definitely claim superiority for method 5. But now, the inclusion of F0 and intensity data within the framework of multiple regression has confirmed that method 5 is consistently more effective. The following numbers illustrate the differences between method 3 and method 5:

For speaker 3, raw rating data (ALL): adj. r^2 = .015,
method-3-normalized rating data (ALL): adj. r^2 = .081,
method-5-normalized rating data (ALL): adj. r^2 = .080,
here showing no superiority of method 5.

For speaker 7, raw rating data (ALL): adj. r^2 = .039,
method-3-normalized rating data (ALL): adj. r^2 = .105,
method-5-normalized rating data (ALL): adj. r^2 = .121,
showing considerable superiority of method 5.

For the rest of the speakers, the differences lie in between, showing superiority of method 5 to some extent or another.

4. Outlook

Method 5 involves mean-based central tendency normalization with respect to one rating and additionally dispersion normalization with respect to all 8 ratings delivered by one listener. (For methodological reasons, our dispersion normalization referred to mean absolute deviation.) The consequences of using sd for reference remain to be investigated, but informal comparisons showed no remarkable differences in the overall picture. More significant is what happens when the 14 rating measures per phonetic syllable are averaged in addition to normalization (data: SMALL, see section 2.4., problem 1), here again illustrated by speakers 3 and 7 and the respective results of the multiple regression analysis:

For speaker 3, raw rating data (SMALL): adj. r^2 = .273,
method-3-normalized rating data (SMALL): adj. r^2 = .290,
method-5-normalized rating data (SMALL): adj. r^2 = .244,
this time showing superiority of method 3.

For speaker 7, raw rating data (SMALL): adj. r^2 = .418,
method-3-normalized rating data (SMALL). adj. r^2 = .424,
method-5-normalized rating data (SMALL): adj. r^2 = .429,
showing only minute superiority of method 5.

The most obvious effect of averaging is that all values are very much higher than before, because the averaging procedure neutralizes inter-listener dispersion altogether. Furthermore, the effort going along with the methods discussed before seems to be unnecessary altogether, because success is absent or poor from this perspective. We conclude that if the goal of a survey is plainly to arrive at one single perception-based datum per phonetic syllable, just averaging the raw rating data is sufficient. If information as carried by inter-listener dispersion is to be taken into account, our results lead to the recommendation of method 5. The fact that this type of information may play an interesting role can only be illustrated here: Figure 7 shows the boxplots of the method-5-normalized rating data corresponding to the boxplots of the raw rating data in Figure 2.

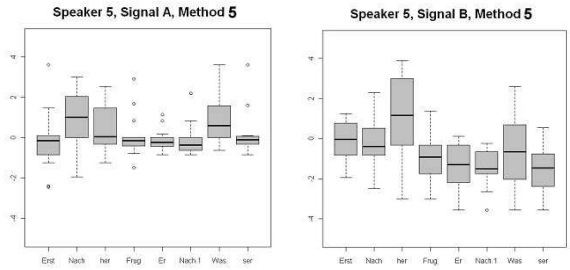

Figure 7: *Boxplots of method-5-normalized prominence ratings (14 per phonetic syllable) concerning two individual readings of the last sentence of stanza 3 (see (4), (5), Figure 2).*

In this perspective, the median may be more informative than the average, and inter-quartiles give valuable hints at how collective tendencies come about. Closer analyses, for example when rhythm phenomena are taken into view with the help of prominence ratings, can account for this type of information more reliably after the raw rating data has been normalized according to method 5 (Figure 2 vs. Figure 7).

5. References

[1] Sappok, C. and Arnold, D., On the Normalization of Syllable Prominence Ratings. In: Proceedings of Speech Prosody 2012, Shanghai, China, 2012.

[2] Fant, G. and Kruckenberg, A., Preliminaries to the study of Swedish prose reading and reading style. STR-QPSR, 2/1989, pp. 1–80, KTH, Stockholm, 1989.

[3] R Development Core Team, R: A language and environment for statistical computing (Version 2.12.2). R Foundation for Statistical Computing, Vienna, Austria. 2011. URL: http://www.R-project.org.

[4] Heine, H., Historisch-kritische Gesamtausgabe der Werke. Bd. 3. Romanzero, Gedichte. 1853 und 1854, Lyrischer Nachlaß (Windfuhr, M., ed), Hamburg, Germany, 1992

[5] Boersma, P. and Weenink, D., Praat: doing phonetics by computer (Version 5.1.31) [Computer program], 2011. URL: http://www.praat.org.

[6] Gibbon, D., SAMPA-D-VMlex Dokumentation V1.0, Bielefeld, Germany, 1995.